Abstract

High performance grid computing is a key enabler of large scale collaborative computational science. With the promise of exascale computing, high performance grid systems are expected to incur electricity bills that grow super-linearly over time. In order to achieve cost effectiveness in these systems, it is essential for the scheduling algorithms to exploit electricity price variations, both in space and time, that are prevalent in the dynamic electricity price markets. Typically, a job submission in the batch queues used in these systems incurs a variable queue waiting time before the resources necessary for its execution become available. In variably-priced electricity markets, the electricity prices fluctuate over discrete intervals of time. Hence, the electricity prices incurred during a job execution will depend on the start and end time of the job.

Our thesis consists of two parts. In the first part, we develop a method to predict the start and end time of a job at each system in the grid. In batch queue systems, similar jobs which arrive during similar system queue and processor states, experience similar queue waiting times. We have developed an adaptive algorithm for the prediction of queue waiting times on a parallel system based on spatial clustering of the history of job submissions at the system. We represent each job as a point in a feature space using the job characteristics, queue state and the state of the compute nodes at the time of job submission. For each incoming job, we use an adaptive distance function, which assigns a real valued distance to each history job submission based on its similarity to the incoming job. Using a spatial clustering algorithm and a simple empirical characterization of the system states, we identify an appropriate prediction model for the job from among standard deviation minimization method, ridge regression and k-weighted average. We have evaluated our adaptive prediction framework using historical production workload traces of many supercomputer systems with varying system and job characteristics, including two Top500 systems. Across workloads, our predictions result in up to 22% reduction in the average absolute error and up to 56% reduction in the percentage prediction errors over existing techniques. To predict the execution time of a job, we use a simple model based on the estimate of job runtime provided by the user at the time of job submission.

In the second part of the thesis, we have developed a metascheduling algorithm that sched-

ules jobs to the individual batch systems of a grid, to reduce both the electricity prices for the systems and response times for the users. We formulate the metascheduling problem as a Minimum Cost Maximum Flow problem and leverage execution period and electricity price predictions to accurately estimate the cost of job execution at a system. The network simplex algorithm is used to minimize the response time and electricity cost of job execution using an appropriate flow network. Using trace based simulation with real and synthetic workload traces, and real electricity price data sets, we demonstrate our approach on two currently operational grids, XSEDE and NorduGrid. Our experimental setup collectively constitute more than 433K processors spread across 58 compute systems in 17 geographically distributed locations. Experiments show that our approach simultaneously optimizes the total electricity cost and the average response time of the grid, without being unfair to users of the local batch systems. Considering that currently operational HPC systems budget millions of dollars for annual operational costs, our approach which can save $167K in annual electricity bills, compared to a baseline strategy, for one of the grids in our test suite with over 76000 cores, is very relevant for reducing grid operational costs in the coming years.